DOCUMENT RESUME

ED 189 113 TM 800 276

AUTHOR Divgi, D. R.
TITLE Dimensionality of Binary Items: Use of a Mixed Model.
PUB DATE Apr 80
NOTE 11p.: Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 8-10, 1980).

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Achievement Tests: Error of Measurement: *Factor Analysis: Factor Structure: *Item Analysis: *Latent Trait Theory: Test Interpretation: Test Items
IDENTIFIERS *Metropolitan Achievement Tests

ABSTRACT

Because it is difficult to ascertain the dimensionality of a test composed of binary items through the use of factor analysis alone, a method is proposed that combines item characteristic curve (ICC) theory with factor analysis. Factor structure of tetrachoric correlations is distorted by non-normal distribution of ability. Item characteristics curve theory cannot demonstrate multidimensionality. Therefore, a model is proposed in which the item score is expressed as the sum of an ICC function representing the general factor, and a residual obeying a linear factor model. The former is calculated using estimated ability and subtracted from the observed item score to obtain the residual. Residuals are factor analyzed by standard methods. An illustration is provided using reading and mathematics items. The results show that the variance due to factors specific to subtests is small compared to error even when subtests differ widely in content. (Author/CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED189113

Dimensionality of Binary Items: Use of a Mixed Model

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
D. Divgi
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

D. R. Divgi

Syracuse University
and
The University of Iowa

ABSTRACT

It is difficult to judge the dimensionality of a test composed of binary items. Factor structure of tetrachoric correlations is distorted by non-normal distribution of ability. Item characteristic curve (ICC) theory cannot demonstrate multidimensionality. Therefore a mixed model is proposed. The item score is expressed as the sum of an ICC function representing the general factor, and a residual obeying a linear factor model. The former is calculated using estimated ability and subtracted from the observed item score to obtain the residual. Residuals are factor analyzed by standard methods. An illustration is provided using reading and mathematics items. The results show that the variance due factors specific to subtests is small compared to error even when subtests differ widely in content.

Paper presented at the 1980 Annual Meeting of the National Council on Measurement in Education, Boston, April 1980.

TM 800276

Dimensionality of Binary Items: Use of a Mixed Model

Introduction

One basic assumption in item characteristic curve (ICC) theory is that all items in the test measure the same latent trait; i.e., that the test is unidimensional. Testing this assumption is a vexing problem. One can argue that if the model fits the data, this provides support for all the underlying assumptions, including unidimensionality. One difficulty is that even when a rigorous test is available, it is valid only in the limit of infinite sample size (Hambleton et al, 1978, p. 489). Moreover, if the model does not fit the data, one does not know whether it is unidimensionality or some other assumption that is violated.

Factor analysis is an attempt to study the dimensionality directly. However, since the variables involved are binary (scored 0 or 1), one must choose a correlation coefficient first. It is easy to prove that the factor structure of phi coefficients depends on the ability level of the group (Lord & Novick, 1968, p. 349). Tetrachoric correlations are not satisfactory either. Lord and Novick (1968, p. 382) have emphasized that presence of a single factor in the matrix of tetrachoric correlations is sufficient but not necessary for inferring that the test is unidimensional. Even when all items measure the same ability, one will not obtain a unifactor matrix of tetrachorics unless the ability is normally distributed. Another source of spurious factors is guessing, which lowers correlations (Carroll, 1945; Lord & Novick, 1968). Hambleton and Traub (1973) studied three tests. One of these was a mathematics test with only 20 items. It was found to require at least two factors for an acceptable factor solution. Another problem with factoring tetrachorics is that it is impossible to obtain factor scores. Tetrachorics are correlations among hypothetical variables. As the variables being factor analyzed are themselves unobservable, there is no possibility of calculating factor scores.

Linear and Nonlinear Models

Consider a factor analytic model with m common factors. The observed value of a variable X_g is given by

$$X_g - \bar{X}_g = \sum_{i=1}^{m} a_{gi} C_i + U_g$$

where C's are common factors, a_{gi} is the loading of variable g on factor i, and U_g is the unique part of the variable, including error. The common and unique factors are usually conceived of as continuous variables. Therefore this model is not suitable for item scores which can only be zero or one (Lord & Novick, 1968, p. 537;

MacDonald & Ahlawat, 1974). Moreover, it is easy to see that the linear relationship assumed in the model is completely unrealistic. By assumption,

$$E\ (U_g \mid c_1, \ldots, c_m) = 0$$

for any g and at any values of the common factors. Therefore

$$E\ (X_g \mid c_1, \ldots, c_m) = \sum_{i=1}^{m} a_{gi} c_i + \bar{X}_g \qquad (1)$$

Consider a test with a single common factor C. The loading a_g determines how strongly the item score depends on the factor, and therefore represents discrimination. Let the loadings be equal for all items. Then, for any two items g and h,

$$E\ (X_g - X_h \mid c) = \bar{X}_g - \bar{X}_h \qquad (2).$$

Thus the model predicts that the difference between expected values is independent of ability. Let item g be easier than item h ($\bar{X}_g > \bar{X}_h$). If an examinee's ability is so high as to guarantee a correct answer on item h [$E\ (\bar{X}_h \mid c) = 1$], Eqn. (2) implies that the expected score on the easier item g exceeds unity, which is impossible.

Logistic and normal ogive models do not suffer from this defect because they are constructed specifically for binary items (Hambleton et al, 1978). Any ICC model postulates a trait or ability θ which is measured by the test. Higher ability implies greater chance of answering any given item correctly. This probability is the expected score on the item, and depends on characteristics of the item such as difficulty and discrimination. The characteristic curve of an item is the functional dependence of expected score on ability. Let us denote it by $P_g(\theta)$. Its form and the number of item parameters vary from one model to another. However the function is always monotonic increasing, has an upper asymptote of unity, and a lower asymptote of zero or more. Therefore ICC models do not lead to anomalies such as Eqn.(2). They do not require normality assumptions as the use of tetrachoric correlations does, and the effect of guessing can be handled by using the three-parameter model. On the other hand these models deal with only a single trait and cannot be used to examine additional factors if any.

A Mixed Model

One can express the item score as the sum of the expected value and a residual:

$$X_g = P_g(\theta) + r_g$$

If the test is in fact unidimensional, r_g is pure error and therefore residuals of

different items are uncorrelated. However, suppose the test consists of two subtests, say Reading and Mathematics, each of which is homogeneous. The subtests measure two abilities which are strongly but not perfectly correlated. When we use the test as a single instrument, what we measure is an average (in some sense) of these abilities (Reckase, 1979). Denote it by θ_a. For persons who are better at reading than at mathematics, θ_a is smaller than the reading ability and $P_g(\theta_a)$ is an underestimate of the expected score if g is any Reading item. Therefore the residual $x_g - P_g(\theta_a)$ will tend to be positive for a Reading item and negative for a Mathematics item. When we consider a population in which some examinees are better at reading and others at mathematics, residuals of two Reading or two Mathematics items will be positively correlated, and residuals for a Reading and a Mathematics item will be negatively correlated. Therefore a factor analysis of correlations will yield a factor representing the difference between Reading and Mathematics. In contrast to item scores, residuals are continuous (ranging from -1 to 1) and therefore much more amenable to analyses based on product-moment correlations and linear models. These considerations suggest that factor analysis of residuals can be used to detect multidimensionality in a test.

The argument can be stated more systematically as follows. Let θ_R and θ_M be the abilities measured by the Reading and Mathematics tests respectively. Let $\theta_a = (\theta_R + \theta_M)/2$ represent the general ability and $\theta_d = (\theta_R - \theta_M)/2$ represent the specific factor, so that $\theta_R = \theta_a + \theta_d$ and $\theta_M = \theta_a - \theta_d$. We assume that the correlation between θ_R and θ_M is high and hence the variance of θ_d is small. Then, for an item g, the probability of a correct answer is

$$P_g(\theta_a + s_g\theta_d) = P_g(\theta_a) + s_g P'_g(\theta_a)\theta_d$$

where s_g is the sign factor, +1 for Reading and -1 for Mathematics items. Then

$$r_g = X_g - P_g(\theta_a) = s_g P'_g(\theta_a)\theta_d + e_g$$

contains not only the random error e_g but also a term proportional to the difference between the abilities. The value of P'_g depends on θ_a, but it is always positive. Its average over the population is analogous to the conventional factor loading.

For the general case, we propose the following mixed model.

$$X_g = P_g(\theta) + r_g; \qquad (3a)$$

$$r_g = \sum_{i=1}^{m} a_{gi} C_i + e_g \,. \qquad (3b)$$

$P_g(\theta)$, given by any ICC model one may choose, represents the dominant general factor, and replaces the first principal factor in conventional factor analysis. (If the test is in fact unidimensional, $P_g(\theta)$ is the true item score, and r_g is completely random.) m is the number of factors remaining after the general one is subtracted, and e_g is random error. This model is offered purely as a compromise combining the realistic ICC of a latent trait model with the capacity of the linear model to study multiple factors. No theoretical basis is claimed. It may be seen as a special form of analysis of residuals familiar in regression theory. The model does not require any assumption about the distribution of ability. The X's are observed item scores and θ can be estimated from item scores by well-established procedures. Thus the estimate $\hat{r}_g = x_g - P_g(\hat{\theta})$ is an observable quantity, and therefore factor scores for i = 1, ..., m can be estimated by the usual methods. The procedure is as follows:

a) Calibrate the ICC model, i.e., estimate the item parameters.

b) Use these parameters to calculate ability estimate $\hat{\theta}$ for each examinee.

c) Use $\hat{\theta}$ to calculate predicted score, i.e., probability of a correct answer, on each item. Subtract this from the actual item score (0 or 1) to obtain the residual $\hat{r}_g$. With an n-item test there will be n residual scores for each person. The sum of the probabilities over items is the estimated score on the first factor.

d) Analyze covariances or correlations of the residuals by any method of factor analysis. Component scores can be calculated or factor scores estimated if desired.

Whether the residuals provide evidence for specific factors can be judged only by subjective criteria. Certainly, one cannot simply ask whether the eigenvalue exceeds unity. Several factors will pass this criterion purely due to chance correlations (Lord & Novick, 1968, p. 382). Even if the ICC model holds rigorously, residuals of the items will have small negative correlations because expected values are calculated using an ability estimate based on those same items. (This effect is clearest with the Rasch model, where the estimate $\hat{\theta}$ is that value which satisfies $\sum_{g=1}^{n} P_g(\hat{\theta}) = \sum_{g=1}^{n} x_g$. Therefore $\sum_g \hat{r}_g = 0$ for each person.) The longer the test, the smaller these correlations, because any single item has less influence on $\hat{\theta}$.

Illustration

The procedure described above was applied to 20 Reading and 20 Mathematics items from the Survey battery of the Metropolitan Achievement Tests (~~The Psychological Corporation~~ Hogan, Farr, Prescott & Balow, 1978). An unselected sample of 1081 examinees was obtained by using every fourth case. The 16th to 35th items were taken from each test, these being in the middle of the tests. Item analysis showed a wide range of p values (.29 to .86) and biserials (.26 to .90) among the items used. A second sample ($N = 1106$) consisted of all examinees with raw Otis-Lennon scores in the top quarter. This is analogous to using only those persons who have been selected for admission into college. Such selection results in a skewed distribution of ability. The same analyses were carried out with both groups, which are identified as "Unselected Sample" and "Top Quarter" in the Tables.

The Rasch model was used for ICC analysis because of its simplicity. It contains only one item parameter, b_g, which represents difficulty of the item. The expected item score is

$$P_g(\theta) = \exp(\theta - b_g)/[1 + \exp(\theta - b_g)].$$

The item parameters were estimated using the approximate procedure of Wright and Douglas (1977) followed by an iterative procedure to obtain maximum likelihood estimates. Three separate calibrations were carried out - Reading, Mathematics, and Total. The difficulty parameters were used to estimate abilities and calculate residuals. (Examinees with zero and perfect scores were ignored.) The correlations among the residuals were factor analyzed.

Results of the Rasch analyses are presented in Table 1. N(Rasch) is the number of usable cases after dropping those with zero or perfect score on the test being analyzed. σ_T^2 denotes the total (i.e. summed over items) variance of scores on the general factor, i.e. of $P_g(\hat{\theta})$. σ_r^2 and σ_X^2 are the corresponding quantities for residuals and observed item scores respectively. σ_X^2 differs very little from $\sigma_T^2 + \sigma_r^2$, which means that residuals are uncorrelated with the general factor, as expected. $(\sigma_T^2 / \sigma_X^2)$ is slightly larger than KR-20 stepped down to a single item, shown in Table 1 as"KR20/item." This is reasonable because KR-20 is a lower bound for true reliability.

There is little evidence of factor structure in the correlation matrices of the residuals. The numbers of eigenvalues above unity (shown as "#λ>1" in Table 1)

are almost or exactly half the numbers of items. The first component explains 8.1% or less of the variance. (The first eigenvalue is shown as "λ_1" in Table 1.) However, if we consider the ratios λ_1/λ_2 and $(\lambda_1-\lambda_2)/(\lambda_2-\lambda_3)$ as indices of the importance of the first component, they are found to be larger for Total than for Reading or Mathematics. The factor matrix provided additional evidence of a common factor among the residuals of the total test. When a single factor was retained, the factor loadings were negative for all Mathematics items and positive for Reading items, the exception being a single Reading item on using the unselected sample. Thus use of the mixed model, with no assumptions beyond those of the Rasch model, demonstrated unidimensionality of the Reading and Mathematics items, and multidimensionality of the total test. Evidence for the latter is admittedly weak; the reason will be discussed below.

Component analyses were performed on the tetrachoric correlations also. Results are shown in Table 2. The Reading and Mathematics tests appear unidimensional (according to the criterion of Lord and Novick, 1968, p. 382) as expected, but so does the total test. Only $(\lambda_1-\lambda_2)/(\lambda_2-\lambda_3)$, by being much smaller for the total test, indicates presence of additional factors. More important, comparison of results from the two samples clearly shows the marked effect of a skewed distribution of ability caused by selection of examinees. When correlations from the Top Quarter sample are used, importance of the first factor appears smaller in every test, according to every criterion. This illustrates the point made by Lord and Novick (1968) that the factor structure of tetrachoric correlations is not a reliable indicator of the dimensionality of a test unless the distribution of ability is at least approximately normal. No such restrictions apply to use of the mixed model, because no distributional assumptions are made. The only major difference between results from the two samples is that the selected sample yields smaller score variances and hence lower reliabilities. In particular, the clearest piece of evidence for presence of subtests in the total test - signs of loadings on the first factor - is the same in both samples. (It seems worth mentioning that analysis of Rasch would probably cost the same as calculation of tetrachoric correlations. With the approximation of Wright and Douglas, 1977, as the initial estimate, calibration of the Rasch model is highly efficient. Analysis of tetrachorics involves calculation of $n(n-1)/2$ correlations by an iterative procedure. For a recent algorithm see Divgi, 1979.)

Conclusion

The results indicate that the mixed model presented here is superior to factor analysis of tetrachoric correlations because it does not require normal distribution of ability. When Reading and Mathematics items were combined to form a single test, factor analysis of the residuals provided some evidence for the presence of distinct subtests, mainly through signs of loadings on the first common factor. The effect of this factor on eigenvalues was barely noticeable. This happened despite the fact that there were only two, clearly different, subtests. The question arises: was this due to a shortcoming in the method or inherent in the nature of the problem?

It seems unlikely that any other method can yield appreciably better results with the same sample size. The reason is this. What we are looking for in the residuals is evidence for one or more common factors in addition to the general ability. Let us estimate the contribution of the additional factor in the items analyzed in this study. In the unselected sample the correlation between Reading and Mathematics was .64 (.30 in the other sample). After correction for attenuation, the correlation between true scores, τ_R and τ_M, in Reading and Mathematics is .79. The general factor may be represented by $(\tau_R + \tau_M)/\sqrt{2}$ and the ~~additional factor~~ specific factor by $(\tau_R - \tau_M)/\sqrt{2}$. Using values of σ_T^2 for Reading and Mathematics, the variance of these quantities are estimated as 1.83 and 0.24 respectively. The latter is only 3.6% of the total error variance 3.14 + 3.52 = 6.66.* This implies an average communality and average correlation of about .035, which is of the same order as random error with a sample size of about 1000. Thus the variance accounted for by the difference between subtests as clearly distinct as Reading and Mathematics is a small fraction of the total, and therefore hard to detect. The same thing is likely to happen with any composite test used in practice.

This is not to say that the distinction between the subtests is negligible. Their corrected correlation of .79 is clearly less than perfect. When items are added up to form subtest scores, reliability increases and the importance of error decreases. Therefore what is barely noticeable at the level of single items can become quite clear at the level of aggregates - provided items which belong to different subtests are identified in advance. If one refuses to make such judgments, analysis must remain at the level of individual items and therefore lack power. The conclusions from any such analysis will depend strongly on subjective criteria such as rules for choosing the number of factors.

---

*Variances in subtests add up to more than those in the total. The reason is that about 10% or more examinees were excluded from the Rasch analysis of Reading because of perfect scores.

Acknowledgments

I am grateful to the Psychological Corporation for providing data for this research.

REFERENCES

Carroll, J. B. The effect of difficulty and chance success on correlations between items or between tests. Psychometrika, 1945, 10, 1-19.

Divgi, D. R. Calculation of the tetrachoric correlation coefficient. Psychometrika, 1979, 44, 169-172.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Hambleton, R. K. & Traub, R. E. Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 1973, 26, 195-211.

Hogan, T. P., Farr, R. C., Prescott, G. A., & Balow, I. H. Metrołopolitan Achievement Tests. New York: The Psychological Corporation, 1978.

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison Wesley, 1968.

McDonald, R. P. & Ahlawat, K. S. Difficulty factors in binary data. British Journal of Mathematical and Statistical Psychology, 1974, 27, 82-99.

Reckase, M. D. Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 1979, 4, 207-230.

Wright, B. D. & Douglas, G. A. Best procedures for sample-free item analysis. Applied Psychological Measurement, 1977, 1, 281-295.

Table I

Test Statistics and Results of Analyses of Residuals

	Unselected Sample			Top Quarter		
Statistic	Reading	Mathematics	Total	Reading	Mathematics	Total
Mean	13.3	11.0	24.3	18.0	14.9	32.8
St. Dev.	5.17	3.89	8.23	2.23	2.55	3.87
KR-20	.885	.753	.896	.687	.532	.685
KR-20/Item	.278	.132	.177	.099	.053	.052
N(Rasch)	992	1077	1081	822	1094	1100
σ_T^2	1.28	0.79	1.78	0.28	0.38	0.49
σ_R^2	3.14	3.52	6.81	1.90	2.85	4.51
σ_X^2	4.40	4.31	8.57	2.16	3.23	4.95
λ_1	1.62	1.49	2.57	1.50	1.48	2.14
$\#\lambda>1$	9	10	17	11	11	19
λ_1/λ_2	1.15	1.03	1.54	1.07	1.07	1.37
$(\lambda_1-\lambda_2)/(\lambda_2-\lambda_3)$	2.05	0.91	4.43	0.75	1.08	8.23

Table II

Results of Component Analyses of Tetrachoric Correlations

	Unselected Sample			Top Quarter		
Statistic	Reading	Mathematics	Total	Reading	Mathematics	Total
λ_1	9.75	5.83	13.76	6.61	4.45	8.46
$\#\lambda>1$	2	5	8	4	6	14
λ_1/λ_2	8.76	4.11	6.37	4.27	2.88	2.88
$(\lambda_1-\lambda_2)/(\lambda_2-\lambda_3)$	53.1	35.3	18.0	23.5	22.0	6.0